МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение высшего образования

«ИРКУТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

(ФГБОУ ВО «ИГУ»)

Институт математики и информационных технологий

Кафедра вычислительной математики и оптимизации

«УТВЕРЖДАЮ»
Директор ИМИТ ИГУ
М. В. Фалалеев
«15» мая 2024 г.

Рабочая программа дисциплины (модуля)

Б1.О.09 Большие данные

Направление подготовки	01.04.02 Прикладная математика и информатика
Направленность (профиль) подготовки	Цифровая бизнес-аналитика
Квалификация выпускника	магистр
Форма обучения	очная

Иркутск 2024 г.

1. ЦЕЛИ И ЗАДАЧИ ДИСЦИПЛИНЫ

Цели: - формирование компетенций специалиста по направлению «Прикладная математика и информатика» в предметной области, связанной с решением задач сбора и анализа огромных объемов структурированной или слабоструктурированной информации, разработке на ее основе моделей данных и извлечении новых знаний.

Задачи:

- ✓ приобретение студентами знаний о технологиях подготовки, хранения, обработки и анализа больших данных;
- ✓ применение статистических и математических методов для анализа больших объемов информации;
- ✓ приобретение практических навыков работы с большими данными с помощью сред на базе языков программирования Python (R);

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП ВО

Учебная дисциплина Б1.О.10 Большие данные относится к обязательной части Блока 1 образовательной программы.
Для изучения данной учебной дисциплины необходимы знания, умения и навыки, формируемые предшествующими дисциплинами:

Б1.О.01 Управление исследовательской и проектной деятельностью
Б1.О.08 Защита информации
Б1.О.09 Базы данных
Б1.О.11 Информационно-коммуникационные технологии и системы
Б1.В.ДВ.02.01 Технологии программирования в эколого-экономических расчетах

Перечень последующих учебных дисциплин, для которых необходимы знания, умения и навыки, формируемые данной учебной дисциплиной:

Б2.В.01 (П) «Научно-исследовательская работа»;
Б2.О.01 (Пд) «Преддипломная практика»;
Б3.01 «Выполнение и защита выпускной квалификационной работы»

3. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Процесс освоения дисциплины направлен на формирование следующих компетенций в соответствии с ФГОС ВО и ОП ВО по направлению подготовки 01.04.02 Прикладная математика и информатика:
ОПК-3 Способен разрабатывать математические модели и проводить их анализ при решении задач в области профессиональной деятельности.

В результате освоения дисциплины обучающийся должен
знать:

- методы системного анализа и математической статистики;
- современных программных средств анализа больших объемов информации;
- принципы организации современного программного обеспечения при работе с большими данными;
- технологии баз данных и их информационном обслуживании при работе с большими объемами информации.

уметь:

− использовать методы системного анализа и математической статистики для решения эколого-экономических задач;
− анализировать и выбирать оптимальные программные средства для анализа больших данных;
− осуществлять ведение базы данных, обработку и анализ данных.

владеть:

− терминологией и основными методами математической статистики; навыками применения статистических методов для обработки и анализа больших объемов информации;
− навыками использования современных информационно-коммуникационных технологий для решения прикладных задач;
− навыками работы с современными программными средствами анализа данных и способностями осуществлять ведение баз данных в рамках поддержки информационного обеспечения решения прикладных задач;

навыками применения современных программных средств анализа больших объемов информации.

4. СОДЕРЖАНИЕ И СТРУКТУРА ДИСЦИПЛИНЫ

Объем дисциплины составляет 5 зачетных ед., 180 час.

Форма промежуточной аттестации: экзамен.

4.1. Содержание дисциплины, структурированное по темам, с указанием видов учебных занятий и отведенного на них количества академических часов

Раздел дисциплины / тема	Сем.	Виды учебной работы				Формы текущего контроля; Формы промежут. аттестации
		Контактная работа преподавателя с обучающимися			Самост. работа	
		Лекции	Лаб. занятия	Практ. занятия		
Введение в большие данные	3	4		2		Презентация
Жизненный цикл анализа больших данных	3	6		4		Реферат
Корреляция и регрессия. Их роль в аналитике больших данных	3	8		4	13	Защита отчета по практической работе
Технологии хранения и обработки больших данных	3	6		2	12	Защита отчета по практической работе
Языки Python и R, стек библиотек анализа данных. Готовые решения анализа данных (Weka и т.д.).	3	8		4	8	Защита отчета по практической работе
Переподготовка данных. Визуализация данных. Понимание данных	3	6		2	4	Защита отчета по практической работе
Парадигма Map Reduce. Ее реализация Hadoop	3	8		4	4	Защита отчета по практической работе
Проблема переобучения. Регуляризация. Нейронные сети. Машина опорных векторов.	3	6		4	4	Защита отчета по практической работе
Научные проблемы в области больших данных	3	8		4		Реферат
Итого (3 семестр):		60		30	45	экз.

4.2. План внеаудиторной самостоятельной работы обучающихся по дисциплине

Раздел дисциплины / тема	Самостоятельная работа обучающихся			Оценочное средство	Учебно-методическое обеспечение самост. работы
	Вид самост. работы	Сроки выполнения	Затраты времени		

Корреляция и регрессия. Их роль в аналитике больших данных (ОПК-3)	1 Работа со специализированными пакетами алгоритмов Python (R). 2. Подготовка практической работы.	К окончанию выполнения практической работы по данной теме	13	Аналитический отчет (markdown)	ОЛ –1,2 ДЛ - 1,2,3 ИР – 11
Технологии хранения и обработки больших данных	1 Проектирование базы данных. 2. Подготовка практической работы.	К окончанию выполнения практической работы по данной теме	12	Аналитический отчет (markdown)	ОЛ –1 ДЛ - 3 ИР – 5,6,7,8,9
Анализ стандартных наборов данных (iris, mtcars и т.д.) при помощи Weka или Orange.	1. Работа со специализированными пакетами алгоритмов R-Studio. 2. Подготовка практической работы.	К окончанию выполнения практической работы по данной теме	8	Аналитический отчет (markdown)	ОЛ -1, ИР – 7,8
Переподготовка данных. Визуализация данных. Понимание данных.	1. Изучение контрольных примеров визуализация стандартных наборов данных при помощи Tableau.. 2. Подготовка практической работы.	К окончанию выполнения практической работы по данной теме	4	Аналитический отчет (markdown)	ОЛ -1, ИР – 5,6, 7,8
Парадигма Map Reduce. Ее реализация Hadoop	1. Исследование и анализ выбранной предметной области. 2. Подготовка практической работы.	К окончанию выполнения практической работы по данной теме	4	Аналитический отчет (markdown)	ОЛ -1,2 ДЛ – 1,2,3 ИР – 9
Проблема переобучения. Регуляризация. Нейронные сети. Машина опорных векторов	1. Регуляризация для метода наименьших квадратов. 2. Реализация нейронной сети и машины опорных векторов. 2. Подготовка проектной работы.	К окончанию выполнения практической работы по данной теме	4	Аналитический отчет (markdown)	ОЛ -1,2 ДЛ – 1,2,3 ИР –2,3,9
Общая трудоемкость самостоятельной работы (час.)			45		
Из них с использованием электронного обучения и дистанционных образовательных технологий (час.)			45		

**ОЛ – основная литература*

*** ДЛ – дополнительная литература*
**** ИР – интернет - ресурсы*

4.3. Содержание учебного материала

Тема 1. Введение в большие данные

Основные определения, термины, задачи анализа больших данных. Понятие Data Mining. Системный анализ и методы его проведения. Методики анализа больших данных

Тема 2. Жизненный цикл анализа больших данных

Создание данных (Data Generation/Data Capture). Создание данных (Data Generation/Data Capture). Создание данных (Data Generation/Data Capture). Использование данных (Data Usage). Публикация данных (Data Publication). Публикация данных (Data Publication). Уничтожение данных (Data Purging). Жизненный цикл метаданных.

Тема 3. Корреляция и регрессионный анализ.

Корреляция и регрессионный анализ. Коэффициент корреляции. Графическое представление. Постановка задачи регрессионного анализа. Линейная регрессия. Метод наименьших квадратов. Их роль в аналитике больших данных.

Тема 4. Технологии хранения и обработки больших данных

Обзор технологий хранения больших данных. Базы данных. Системы управления базами данных. Модели данных. Подготовка исходных данных для анализа: первичная обработка и визуализация имеющихся данных. Базы данных NoSQL.

Тема 5. Языки Python и R. Синтаксис языка R, основные типы данных

Роль языков программирования Python и R в аналитике больших данных. Необходимый набор библиотек. Готовые решения анализа данных и их роль в области больших данных.

Тема 6. Подготовка данных. Визуализация данных. Понимание данных.

Методы предварительной подготовки данных. Инструменты и методы визуализации данных.

Тема 7. Парадигма Map Reduce. Ее реализация Hadoop

Парадигма Map Reduce. Роль Map Reduce в аналитике больших данных. Оператор Map. Лямбда-архитектура.

Тема 8. Проблема переобучения. Регуляризация. Нейронные сети. Машина опорных векторов

Проблема переобучения и регуляризация. Разбор алгоритма нейронных сетей. Разбор алгоритма SVM

Тема 9. Научные проблемы в области больших данных

Тенденции развития техники больших данных. Консолидация данных. Визуализация. Классификация. Кластеризация. Регрессионный анализ. Анализ ассотиативных правил. Нейронные сети. Технологии и инструменты больших данных. Новые SQL базы данных. Apache Hadoop. Storm – система потоковой обработки. Язык программирования R. Аналитика больших данных как корпоративный проект.

4.3.1. Перечень семинарских, практических занятий и лабораторных работ

Тема занятия	Всего часов	Оценочные средства	Формируемые компетенции
Введение в большие данные	2	Презентация (LaTeX)	ОПК-3
Жизненный цикл анализа больших данных	4	Реферат (LaTeX)	ОПК-3
Примеры использования корреляции и регрессионного анализа в области больших данных.	4	Аналитический отчет (markdown)	ОПК-3
Технологии хранения и обработки больших данных.	2	Проектный отчет (markdown)	ОПК-3
R Programming. Анализ стандартных наборов данных (iris, mtcars и т.д.) при помощи Weka или Orange.	4	Аналитический отчет (markdown)	ОПК-3
Introduction to Data Science.	2	Аналитический отчет (markdown)	ОПК-3
Подсчет количества слов, вычисление индекса TFIDF, реализация алгоритма k-means в рамках парадигмы Map Reduce с использованием Hadoop.	4	Аналитический отчет (markdown)	ОПК-3
Регуляризация для метода наименьших квадратов. Нейронная сеть. Машина опорных векторов.	4	Аналитический отчет (markdown)	ОПК-3
Научные проблемы в области больших данных	4	Реферат (LaTeX)	ОПК-3

4.3.2. Перечень тем (вопросов), выносимых на самостоятельное изучение студентами в рамках самостоятельной работы

Тема	Задание	Формируемые компетенции
Функционал библиотеки NumPy	1. Изучение основных модулей и функций библиотеки. 2. Подготовка опорного конспекта.	ОПК-3
Библиотека Pandas	1. Изучение основных модулей и функций библиотеки. 2. Подготовка опорного конспекта.	ОПК-3
Библиотека Matplotlib	1. Изучение основных модулей и функций библиотеки. 2. Подготовка опорного конспекта.	ОПК-3
Библиотека SciPy.	1. Изучение основных модулей и функций библиотеки. 2. Подготовка опорного конспекта.	ОПК-3
Корреляционный анализ	1. Изучение основных алгоритмов. 2. Реализация контрольного примера с помощью среды Jupyter Notebook (Python).	ОПК-3
Регрессионный анализ	1. Изучение основных алгоритмов. 2. Реализация контрольного примера с помощью среды RStudio.	ОПК-3
Моделирование динамики популяций	1. Изучение численных алгоритмов для решения задач с дифференциальными уравнениями. 2. Реализация контрольного примера с помощью среды Jupyter Notebook (Python).	ОПК-3

4.4. Методические указания по организации самостоятельной работы студентов

Самостоятельная работа студентов всех форм и видов обучения является одним из обязательных видов образовательной деятельности, обеспечивающей реализацию требований Федеральных государственных стандартов высшего образования. Согласно требованиям нормативных документов самостоятельная работа студентов является обязательным компонентом образовательного процесса, так как она обеспечивает закрепление получаемых на лекционных занятиях знаний путем приобретения навыков осмысления и расширения их содержания, навыков решения актуальных проблем формирования общекультурных и профессиональных компетенций, научно-исследовательской деятельности, подготовки к семинарам, лабораторным работам, сдаче зачетов и экзаменов. Самостоятельная работа студентов представляет собой совокупность аудиторных и внеаудиторных занятий и работ. Самостоятельная работа в рамках образовательного процесса в вузе решает следующие задачи:
– закрепление и расширение знаний, умений, полученных студентами во время аудиторных и внеаудиторных занятий, превращение их в стереотипы умственной и физической деятельности;
– приобретение дополнительных знаний и навыков по дисциплинам учебного плана;
– формирование и развитие знаний и навыков, связанных с научно-исследовательской деятельностью;
– развитие ориентации и установки на качественное освоение образовательной программы;
– развитие навыков самоорганизации;
– формирование самостоятельности мышления, способности к саморазвитию, самосовершенствованию и самореализации;
– выработка навыков эффективной самостоятельной профессиональной теоретической, практической и учебно-исследовательской деятельности.

Подготовка к лекции. Качество освоения содержания конкретной дисциплины прямо зависит от того, насколько студент сам, без внешнего принуждения формирует у себя установку на получение на лекциях новых знаний, дополняющих уже имеющиеся по данной дисциплине. Время на подготовку студентов к двухчасовой лекции по нормативам составляет не менее 0,2 часа.

Подготовка к практическому занятию. Подготовка к практическому занятию включает следующие элементы самостоятельной деятельности: четкое представление цели и задач его проведения; выделение навыков умственной, аналитической, научной деятельности, которые станут результатом предстоящей работы. Выработка навыков осуществляется с помощью получения новой информации об изучаемых процессах и с помощью знания о том, в какой степени в данное время студент владеет методами исследовательской деятельности, которыми он станет пользоваться на практическом занятии. Подготовка к практическому занятию нередко требует подбора материала, данных и специальных источников, с которыми предстоит учебная работа. Студенты должны дома подготовить к занятию 3–4 примера формулировки темы исследования, представленного в монографиях, научных статьях, отчетах. Затем они самостоятельно осуществляют поиск соответствующих источников, определяют актуальность конкретного исследования процессов и явлений, выделяют основные способы доказательства авторами научных работ ценности того, чем они занимаются. В ходе самого практического занятия студенты сначала представляют найденные ими варианты формулировки актуальности исследования, обсуждают их и обосновывают свое мнение о наилучшем варианте. Время на подготовку к практическому занятию по нормативам составляет не менее 0,2 часа.

Подготовка к семинарскому занятию. Самостоятельная подготовка к семинару направлена: на развитие способности к чтению научной и иной литературы; на поиск дополнительной информации, позволяющей глубже разобраться в некоторых вопросах; на выделение при работе с разными источниками необходимой информации, которая требуется для полного ответа на вопросы плана семинарского занятия; на выработку умения правильно выписывать высказывания авторов из имеющихся источников информации, оформлять их по библиографическим нормам; на развитие умения осуществлять анализ выбранных источников информации; на подготовку собственного выступления по обсуждаемым вопросам; на формирование навыка оперативного реагирования на разные мнения, которые могут возникать при обсуждении тех или иных научных проблем. Время на подготовку к семинару по нормативам составляет не менее 0,2 часа.

Подготовка к коллоквиуму. Коллоквиум представляет собой коллективное обсуждение раздела дисциплины на основе самостоятельного изучения этого раздела студентами. Подготовка к данному виду учебных занятий осуществляется в следующем порядке. Преподаватель дает список вопросов, ответы на которые следует получить при изучении определенного перечня научных источников. Студентам во внеаудиторное время необходимо прочитать специальную литературу, выписать из нее ответы на вопросы, которые будут обсуждаться на коллоквиуме, мысленно сформулировать свое мнение по каждому из вопросов, которое они выскажут на занятии. Время на подготовку к коллоквиуму по нормативам составляет не менее 0,2 часа.

Подготовка к контрольной работе. Контрольная работа назначается после изучения определенного раздела (разделов) дисциплины и представляет собой совокупность развернутых письменных ответов студентов на вопросы, которые они заранее получают от преподавателя. Самостоятельная подготовка к контрольной работе включает в себя: — изучение конспектов лекций, раскрывающих материал, знание которого проверяется контрольной работой; повторение учебного материала, полученного при подготовке к семинарским, практическим занятиям и во время их проведения; изучение дополнительной литературы, в которой конкретизируется содержание проверяемых знаний; составление в мысленной форме ответов на поставленные в контрольной работе вопросы; формирование психологической установки на успешное выполнение всех заданий. Время на подготовку к контрольной работе по нормативам составляет 2 часа.

Подготовка к зачету. Самостоятельная подготовка к зачету должна осуществляться в течение всего семестра. Подготовка включает следующие действия: перечитать все лекции, а также материалы, которые готовились к семинарским и практическим занятиям в течение семестра, соотнести эту информацию с вопросами, которые даны к зачету, если информации недостаточно, ответы находят в предложенной преподавателем литературе. Рекомендуется делать краткие записи. Время на подготовку к зачету по нормативам составляет не менее 4 часов.

Подготовка к экзамену. Самостоятельная подготовка к экзамену схожа с подготовкой к зачету, особенно если он дифференцированный. Но объем учебного материала, который нужно восстановить в памяти к экзамену, вновь осмыслить и понять, значительно больше, поэтому требуется больше времени и умственных усилий. Важно сформировать целостное представление о содержании ответа на каждый вопрос, что предполагает знание разных научных трактовок сущности того или иного явления, процесса, умение раскрывать факторы, определяющие их противоречивость, знание имен ученых, изучавших обсуждаемую проблему. Необходимо также привести информацию о материалах эмпирических исследований, что указывает на всестороннюю подготовку студента к экзамену. Время на подготовку к экзамену по нормативам составляет 36 часов для бакалавров.

В ФБГОУ ВО «ИГУ» организация самостоятельной работы студентов регламентируется Положением о самостоятельной работе студентов, принятым Ученым советом ИГУ 22 июня 2012 г.

5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

а) основная литература:

1. Макшанов, А. В. Большие данные. Big Data / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь. — 2-е изд., стер. — Санкт-Петербург : Лань, 2022. — 188 с. — ISBN 978-5-8114-9690-7. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/198599 (дата обращения: 27.05.2022). — Режим доступа: для авториз. пользователей.
2. Миркин, Б. Г. Введение в анализ данных : учебник и практикум / Б. Г. Миркин. — Москва : Издательство Юрайт, 2020. — 174 с. — (Высшее образование). — ISBN 978-5-9916-5009-0. — Текст : электронный // ЭБС Юрайт [сайт]. — URL: http://www.biblio-online.ru/bcode/450262 (дата обращения: 21.06.2020).

б) дополнительная литература:

1. Федоров, Д. Ю. Программирование на языке высокого уровня Python : учебное пособие для вузов / Д. Ю. Федоров. — 3-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2022. — 210 с. — (Высшее образование). — ISBN 978-5-534-14638-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://www.urait.ru/bcode/492920 (дата обращения: 30.05.2022).
2. Воронов, М. В. Системы искусственного интеллекта: учебник и практикум для вузов / М. В. Воронов, В. И. Пименов, И. А. Небаев. — Москва : Издательство Юрайт, 2022. — 256 с. — (Высшее образование). — ISBN 978-5-534-14916-6. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://www.urait.ru/bcode/485440 (дата обращения: 30.05.2022).
3. Гниденко, И. Г. Технологии и методы программирования : учебное пособие для вузов / И. Г. Гниденко, Ф. Ф. Павлов, Д. Ю. Федоров. — Москва : Издательство Юрайт, 2022. — 235 с. — (Высшее образование). — ISBN 978-5-534-02816-4. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://www.urait.ru/bcode/489920 (дата обращения: 30.05.2022).

в) базы данных, информационно-справочные и поисковые системы:

1. Научная электронная библиотека «ELIBRARY.RU» [Электронный ресурс] : сайт. – Режим доступа: http://elibrary.ru/defaultx.asp
2. Открытая электронная база ресурсов и исследований «Университетская информационная система РОССИЯ» [Электронный ресурс] : сайт. – Режим доступа: http://uisrussia.msu.ru
3. Государственная информационная система «Национальная электронная библиотека» [Электронный ресурс] : сайт. – Режим доступа: http://нэб.рф
4. В соответствии с п. 4.3.4. ФГОС ВО, обучающимся в течение всего периода обучения обеспечен неограниченный доступ (удаленный доступ) к электронно-библиотечным системам:
    - ЭБС «Издательство Лань». ООО «Издательство Лань». Контракт № 92 от 12.11.2018 г. Акт от 14.11 2018 г.

- ЭБС ЭЧЗ «Библиотех». Государственный контракт № 019 от 22.02.2011 г. ООО «Библиотех». Лицензионное соглашение № 31 от 22.02.2011 г. Адрес доступа: https://isu.bibliotech.ru/ Срок действия: с 22.11.2011 г. бессрочный.
- ЭБС «Национальный цифровой ресурс «Руконт». ЦКБ «Бибком». Контракт № 91 от 12.11.2018 г. Акт от 14.11.2018 г..
- ЭБС «Айбукс.ру/ibooks.ru». ООО «Айбукс». Контракт № 90 от 12.11.2018 г. Акт № 54 от 14.11.2018 г.
- Электронно-библиотечная система «ЭБС Юрайт». ООО «Электронное издательство Юрайт». Контракт № 70 от 04.10.2018 г.

г) интернет-ресурсы

1. http://wombat.org.ua/AByteOfPython/AByteofPythonRussian-2.01.pdf - это свободная книга по программированию на языке Python.
2. https://www.python.org/ - язык программирования Python
3. Воронцов К.В. Математические методы обучения по прецедентам http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf
4. Математические методы распознавания образов Автор: Л.М. Местецкий (Интернет университет высоких технологий)
5. http://www.intuit.ru/department/graphics/imageproc/4/1.html
6. Онлайн курс Machine learning https://www.coursera.org/course/ml
7. Онлайн курс Big Data Overview https://education.emc.com/academicalliance/elearning/Big_Data_Overview/index.htm
8. Онлайн курс R progamming https://www.coursera.org/course/rprog
9. Онлайн курс Introduction to Data Science https://www.coursera.org/course/datasci
10. Онлайн курс «Введение в аналитику больших массивов данных» http://bit.ly/IntuitBDA.
11. Учебник по статистическому обучению http://statweb.stanford.edu/~tibs/ElemStatLearn/

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Учебно-лабораторное оборудование

ЭТОТ РАЗДЕЛ НЕ ЗАПОЛНЯТЬ

6.2. Программное обеспечение

1. Язык программирования Python 3.7. и выше (открытое программное обеспечение).
2. Язык программирования R 3.x и выше (открытое программное обеспечение).
3. R Studio Desktop 1.3.x и выше (открытое программное обеспечение).
4. Jupyter Notebook (открытое программное обеспечение).
5. Jupyter Lab (открытое программное обеспечение).
6. VisualStudioCode (Microsoft, открытое программное обеспечение)
7. Офисный пакет Microsoft Office Project Professional 2019 (лицензия ИГУ для образовательных учреждений).
8. Редакционно-издательская система MikTeX (открытое программное обеспечение).

7. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

7.1. Оценочные средства текущего контроля

Вид контроля	Контролируемые темы	Контролируемые компетенции
Презентация	Введение в большие данные	ОПК-3
Реферат	Жизненный цикл анализа больших данных Научные проблемы в области больших данных	ОПК-3
Аналитический отчет	Примеры использования корреляции и регрессионного анализа в области больших данных. Технологии хранения и обработки больших данных. R Programming. Анализ стандартных наборов данных (iris, mtcars и т.д.) при помощи Weka или Orange. Introduction to Data Science. Подсчет количества слов, вычисление индекса TFIDF, реализация алгоритма k-means в рамках парадигмы Map Reduce с использованием Hadoop. Регуляризация для метода наименьших квадратов. Нейронная сеть. Машина опорных векторов.	ОПК-3

Примеры оценочных средств текущего контроля

1. Темы исследовательской работы

1. Анализ мнений и сентиментов в текстах Параллельные алгоритмы для анализа данных (GPU).
2. Streaming algorithms.
3. Исследование и визуализация структры семантических сетей.
4. Исследование и визуализация структры политических партий.
5. Исследование и визуализация структры лингвистических сетей.
6. Исследование и визуализация структры финансовых сетей.
7. Исследование и визуализация структры инвестиционных сообществ.
8. Исследование и визуализация взаимодействие сотрудников в организациях
9. Постановка и проведение экспериментов в социальных сетях.
10. Ассоциативные правила. Поиск ассоциативных правил.
11. Кластеризация. Алгоритм кластеризации k-means.
12. Классификация с помощью нейросети.

2. Тесты (пятиминутки) (выбрать правильный ответ, или указать число)

1. О соотношении аналоговой и цифровой информации:
1. Большинство данных в мире в 2011 году содержалось:
i. В цифровом виде
ii. В аналоговом виде
2. В каком веке произошёл перевес объёмов накопленных человечеством данных в сторону цифровых?
3. Объём накопленных человечеством цифровых данных на 2012 год измеряется:
i. Петабайтами
ii. Зеттабайтами
iii. Экзабайтами
iv. Йоттабайтами
4. Сколько Петабайт в Зеттабайте?

2. История больших данных
1. Укажите фактор, способствовавший появлению тренда больших данных
i. Маркетинговые кампании крупных корпораций
ii. Снижение издержек на хранение данных
iii. Появление новых технологий обработки потоковых данных
iv. Выпуск баз данных с обработкой данных в памяти
2. Какие вероятные разочарования тренда больших данных?
i. Из-за угрозы безопасности личной жизни (privacy) граждан будут усложнены процедуры сбора данных, что приведёт к падению ценности больших данных.
3. Отметьте значимые события, повлиявшие на формирование тренда больших данных:
i. Разработка Hadoop
ii. Изобретение принципа MapReduce
iii. Разработка языка Pyhton
iv. Победа Deepblue в матче с Г.Каспаровым.
3. Определение больших данных:
1. Выберите верный ответ
i. Большие данные – это обработка или хранение более 1 Тб информации.
ii. Проблема больших данных – это такая проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна.
iii. Большие данные – это огромная PR-акция крупных вендоров и не более того.
iv. Большие данные – это явление, когда цифровые данные наиболее полно представляют изучаемый объект.
2. Выберите неверный ответ:
i. Большие данные – это данные объёма свыше 1 Тб
ii. Проблема больших данных – это проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна.
iii. Большие данные – это тренд в области ИТ, подогреваемый маркетинговыми кампаниями крупных вендоров.
iv. Большие данные как правило не структурированы.
3. Отметьте те из вариантов, в которых данные структурированы:
i. Данные о продажах компании, представленные в виде помесячных отчётов в формате MS Word.
ii. Таблица с ежедневными показаниями температуры помещения за год в файле формата csv.
iii. Текст педагогической поэмы А.С. Макаренко, представленный в формате PDF.
iv. Библиотека фильмов, представленных в формате mpeg4 на одном жестком диске.
4. Характеристики Big Data:
1. Перечислите четыре основных характеристики Big Data:
i. Virtualization, Volume, Variability, Vehicle
ii. Variety, Velocity, Volume, Value
iii. Verification, Volume, Velocity, Visualization
iv. Video, Value, Variety, Volume
2. Выберите неверное высказывание:
i. Большие объёмы данных приводят к слабой их структуризации, поэтому появляется такое разнообразие данных.

ii. Увеличившаяся производительность телекоммуникационных каналов привела к росту объёмов передаваемой информации.
iii. Удешевление систем хранения на единицу информации привело к росту рынка больших данных.
iv. Большое разнообразие источников данных
3. Отметьте неверное понимание Variety в контексте характеристик Big Data:
i. Высокая скорость генерирования данных.
ii. Разные типы данных в колонках таблиц реляционных СУБД.
iii. Разнообразие отраслей, являющихся источниками данных.
iv. Разнообразие типов данных, включающих в себя структурированные, полуструктурированные и неструктурированные.
5. Принцип MapReduce
1. Принцип MapReduce состоит в том, чтобы
i. Производить вычисления на узлах, где информация изначально была сохранена
ii. Использовать вычислительные мощности систем хранения
iii. Использовать функциональное программирование для решения задач массивно-параллельной обработки
2. Выберите одно неверное высказывание про MapReduce:
i. Интерфейс для массово-параллельной обработки данных, где вычисления производятся на узлах, где информация изначально была сохранена
ii. MapReduce – это две операции: распределения и сборки данных
iii. MapReduce был придуман разработчиками Hadoop
iv. MapReduce был анонсирован разработчиками Google
3. Каков теоретический прирост производительности при подсчёте числа слов в тексте при работе MapReduce при переходе от одного узла к двум?
6. Технологии хранения
1. Какие из следующих технологий СУБД не используют принцип MapReduce
i. Hadoop
ii. Cassandra
iii. HDInsight
iv. Redis
2. Какие СУБД полностью полагаются на оперативную память при хранении информации:
i. Oracle Exalytics
ii. SAP HANA
iii. BigTable
iv. HBase
3. В чём преимущество колоночно-ориентированных СУБД?
i. Они позволяют выполнять более сложные SQL-запросы по сравнению с реляционными СУБД
ii. Они позволяют динамически дополнять содержание записей новыми полями
iii. Они имеют более гибкие возможности аналитики.
iv. Они позволяют эффективно делать межколоночные сравнения.
7. «Песочница» в аналитическом процессе
1. Для чего аналитику необходима «песочница»?
i. Для высокопроизводительной аналитики за счёт использования оперативной памяти и inDB операций.
ii. Для хранения всех полученных от заказчика данных.

iii. Для построения отчётов о результатах анализа
iv. Для снижения затрат, связанных с репликацией данных
2. Какие из следующих средств разумно использовать для анализа данных, представленных единственным csv-файлом размера более 100Гб:
i. Hadoop
ii. Data Warehouse
iii. «Песочница»
iv. Python
3. Выберите верное утверждение:
i. Data Warehouse создаются для проверки гипотез при анализе больших данных.
ii. «Песочница» используется для снижения нагрузки на основной Data Warehouse.
iii. Каждый Data Warehouse должен содержать «песочницу».
iv. «Песочница» необходима для любого процесса аналитики.
8. CRISP-DM
1. Расставьте последовательность этапов проекта аналитики в соответствии с CRISP-DM.
i. Понимание бизнеса (Business understanding)
ii. Понимание данных (Data Understanding)
iii. Подготовка данных (Data Preparation)
iv. Моделирование (Modeling)
v. Оценка (Evaluation)
vi. Внедрение (Deployment)
2. На каком из этапов процесса CRISP-DM происходит проверка гипотез?
i. Понимание бизнеса (Business understanding)
ii. Понимание данных (Data Understanding)
iii. Моделирование (Modeling)
iv. Оценка (Evaluation)
3. Вы являетесь владельцем и аналитиком в компании из 10 человек, в которой требуется проанализировать продажи за 1 год (1 млн. продаж). Какие из этапов CRISP-DM можно опустить:
i. Понимание бизнеса (Business understanding)
ii. Подготовка данных (Data Preparation)
iii. Моделирование (Modeling)
iv. Оценка (Evaluation)
9. Hadoop
1. Пример благоразумного использования Hadoop
i. Анализ 10 Гб данных.
ii. Ежедневное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт).
iii. Посекундное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт).
iv. Построение графика пульса пациента в реальном времени.
2. Начиная с каких размеров данных обоснованно применение кластера Hadoop для хранения данных?
i. 100Гб
ii. 1Тб
iii. 100Тб
iv. 1Пб

3. Hadoop – это:
i. Набор утилит, и программный каркас для выполнения распределённых программ, работающих на кластерах.
ii. Распределённая СУБД, позволяющая обрабатывать большие данные.
iii. Язык выполнения заданий в парадигме MapReduce.
iv. Распределённая файловая система, предназначенная для хранения файлов большого объёма.

Например:
Демонстрационный вариант контрольной работы №1 (№2, №3)
Демонстрационный вариант теста №1 (№2, №3)
Вопросы для собеседования №1 (№2, №3)
Вопросы для коллоквиума №1 (№2, №3)
Темы рефератов и др.

7.2. Оценочные средства для промежуточной аттестации

Список вопросов для промежуточной аттестации:

Примерный перечень вопросов:

1. Понятие Большие данные. Роль цифровой информации в 21 веке
2. Виды массивов данных.
3. Корреляция и регрессионный анализ. Коэффициент корреляции. Графическое представление.
4. Постановка задачи регрессионного анализа. Линейная регрессия. Метод наименьших квадратов. Привести примеры использования регрессионного анализа.
5. Классификация. Признаковое описание объекта и таблица объектсвойства. Постановка задачи. Отличия задачи классификации от задачи регрессии.
6. Определение модели и алгоритма. Процесс обучения. Проблема переобучения. Р
7. Регуляризация. Cross validation. Привести примеры использования алгоритмов классификации.
8. Кластеризация. Метрики. Матрица парных расстояний. Постановка задачи кластеризации. Отличие от задачи классификации.
9. Ассоциативные правила. Определение. Достоверность и поддержка. Отличия построения ассоциативного правила от решающего правила задачи классификации.
10. Парадигма Map Reduce. Описать принцип работы. Нарисовать диаграмму. Перечислить слабые и сильные стороны. Обозначить области применимости.
11. Визуализация. Дать определение визуализации. Показать важность визуализации в аналитике больших данных. Привести примеры использования визуализации.
12. «Жизненный цикл» проекта по аналитике больших данных.
13. Типовая архитектура проекта в области больших данных. Перечислить используемые технологии, указать степень вовлеченности каждой из технологий на каждом этапе работы над проектом.
14. Современные научные проблемы больших данных. Показать значимость проблем, актуальность, связь с областями математики и инженерии.
15. Определение больших данных, ключевые характеристики. Примеры задач больших данных. Основные виды данных

16. Роль аналитика по данным (Data Scientist). Ключевые компетенции аналитика. Отличия BI от Data Science.
17. Использование модели множественной линейной регрессии для прогнозирования экономических показателей.
18. Доверительные интервалы для зависимой переменной.
19. Сглаживание временных рядов. Динамические модели с распределенными лагами.
20. Стационарные временные ряды. Тестирование стационарности.
21. Коинтеграция. Анализ временных рядов.
22. Адаптивные и мультипликативные методы прогнозирования. Экспоненциальное сглаживание.
23. Авторегрессионные модели. Модели скользящего среднего.
24. Интегрированные процессы. Идентификация авторегрессионной модели скользящего среднего.
25. Прогнозирование с моделями временных рядов. Доверительные интервалы прогноза.
26. Предсказание и прогнозирование социально-экономических прогнозов.
27. Дисперсионный анализ влияния качественных факторов. Ранговые методы.
28. Факторный анализ. Метод главных факторов.
29. Многомерное шкалирование. Классическая модель многомерного шкалирования.
30. Неметрические методы. Кластерный анализ. Дискриминантный анализ.
31. Многомерный статистический анализ

Примеры заданий:
Тесты и задания в ЭИОС ИГУ на сайте https://educa.isu.ru/

Разработчик: Кедрин В.С., к.т.н., доцент, доцент